A Systematic Comparison of Different HMM Designs for Emotion Recognition from Acted and Spontaneous Speech

Johannes Wagner, Thurid Vogt, and Elisabeth André

Multimedia concepts and applications
Augsburg University, Germany
johannes.wagner@student.uni-augsburg.de,
{vogt,andre}@informatik.uni-augsburg.de

Abstract. In this work we elaborate the use of hidden Markov models (HMMs) for speech emotion recognition as a dynamic alternative to static modelling approaches. Since previous work on this field does not yet define a clear line which HMM design should be prioritised for this task, we run a systematic analysis of different HMM configurations. Furthermore, experiments are carried out on an acted and a spontaneous emotions corpus, since little is known about the suitability of HMMs for spontaneous speech. Additionally, we consider two different segmentation levels, namely words and utterances. Results are compared with the outcome of a support vector machine classifier trained on global statistics features. While for both databases similar performance was observed on utterance level, the HMM-based approach outperformed static classification on word level. However, setting up general guidelines which kind of models are best suited appeared to be rather difficult.

1 Introduction

For the recognition of emotions from speech, many feature extraction strategies and a number of classification approaches have been explored. These have been mainly static modelling approaches that compute global statistics of relevant features over an "emotion unit", e. g. a word or an utterance [14,2]. However, the temporal structure within the expression of emotions becomes largely lost by this kind of modeling, though it has been noted as an important feature type [4]. In contrast, hidden Markov models (HMMs) offer a dynamic modelling approach which provides a better consideration of the temporal structure but up to now has been used relatively scarcely (e. g. [8,13,5]). This may partly be due to the set of parameters of an HMM (model topology, number of states, output probabilities) which is crucial to its performance though an optimal configuration for emotion recognition is still disputed. In this paper, we address the lack of a systematic analysis of the suitability of different HMM configurations and examine all combinations of the most common settings of the individual parameters.

Furthermore, HMMs have so far been tested mainly on acted emotions or on small spontaneous emotion databases with few speakers. We explore their

use on an acted emotions corpus (BERLIN [3]) and on a spontaneous emotions corpus of considerable size (AIBO [1]) to see whether HMMs are also suitable for spontaneous speech and if differences exist in good model parameters for acted and spontaneous emotion.

In the following we will first introduce HMMs as a modelling technique for emotion recognition, discuss some previous work on this topic and elaborate on our goals. After describing our experimental setting in terms of databases, features and HMM modelling environment, we will present our results along with a thorough interpretation of the findings.

2 HMMs in Speech Emotion Recognition

2.1 Static Versus Dynamic Modelling

A main characteristic of human speech is its dynamic structure. In classification tasks the sampled speech signal is represented as a time series of observations, usually multi-dimensional vectors of relevant features. The length of a single observation is usually around 10–25 ms and as long as the observations are kept in the original sequence, the dynamic information within a segment is captured.

Since many classification methods only handle single observations, a common strategy is the use of global statistics instead of the sequence itself. Popular methods requiring this intermediate step are support vector machines, neural networks or Naïve Bayes, referred to as static or discriminative classifiers. Although experiments set up on these methods have shown reasonable results, as a drawback the temporal structure of the sequences is discarded and consequently also affective information incorporated within the temporal activity of speech.

As an alternative, dynamic modelling methods exist which do not suffer from this drawback. Such a method is discussed within the scope of this work: the so called hidden Markov models (HMMs), which are capable of processing sequences with dynamic length. An HMM is a stochastic finite automaton, where the probability to pass to the next state only depends on the previous state. Additionally, each state produces an output with a certain probability. Only this output can be observed, the under-lying state sequence is "hidden" and has to be inferred from the observations. Thus, an HMM is characterised by the transition probabilities between its states which determine the connectivity of the model, the output probabilities of the states which are usually mixtures of Gaussian distributions, and the number of states in the model. In order to build an HMM classifier, transition and output probabilites are estimated from a training set of data instances.

Transferred to emotion recognition, the output is the observed sequence of feature vectors and the state sequence represents the emotion to be recognised. Obviously, temporal changes in the features can be captured well by this kind of classifier and this is one of the reasons why HMMs are an established modelling technique for automatic speech recognition. Common topologies there have been forward directed networks with no or only short jumps. For emotion recognition, also topologies with backward jumps and more connectivity have been considered (see Fig. 1). Still, it has not yet been systematically investigated which

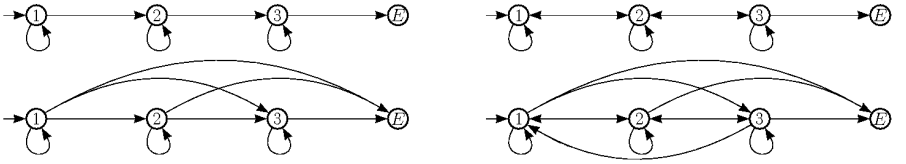

Fig. 1. Different topologies of a 3-state HMM. Top: linear model with only forward connections and model with additional backward connections. Bottom: left-right and fully connected (ergodic) model.

kind of networks are best suited to model the emotional cues and why. Finding reasonable answers to these questions is of main interest to this work.

2.2 Previous Work

Kwon *et al.* [8] used continuous HMMs with left-right topology and up to 5 states to model on word level a neutral and 3 stress styles. They report an average recognition accuracy of 70.1%, which was superior to the performance of a SVM classifier (67.1%). For a second database containing short commands or greetings with several words in 5 basic emotions an average result of 40.8% was achieved, this time inferior to the SVM classifier (42.3%). Based on this observation the authors suggest the use of HMM-based classifiers in applications with short utterances, but discriminative classifiers in case of variable-length utterances. They also report that performance of the HMM-based classifier was improved mostly by increasing the number of states.

Performance of global statistics versus instantaneous features was also addressed by Schuller *et al.* [13]. Utterances collected from 5 speakers in 7 emotional states — acted and spontaneous — were modelled using HMMs with up to 64 states and up to 4 mixtures per state. Again a left-right topology was used with an additional jump limit of two states at most. Performance was compared to the outcome of a Gaussian mixture model (GMM) trained on global statistics of the utterances. While the latter achieved an average recognition accuracy of 86.8%, classification with instantaneous features reached only 77.8%. As a possible reason for this difference the authors quote the elimination of unvoiced parts in the utterances, leading to a loss of temporal information on durations of voiced sounds. Again, adding more states generally improved HMM-based classification. A similar approach is also reported by Jiang *et al.* [6], with the difference that HMM- and GMM-likelihoods were combined to form a final decision.

Beside a single HMM with five states modelled with unimodal Gaussian densities, Fernandez *et al.* [5] also tested several variants such as autoregressive HMMs or hidden Markov decision trees. Experiments were based on utterances recorded from 4 subjects in a driving simulator under different stress situations. Best recognition rates are reported with a mixture of HMMs consisting of several single networks which were trained each on a different cluster of the data. In a subject-dependent task a mean recognition rate of 61.20% for the four stress

levels was achieved. In a second approach based on global statistics, a support vector machine and a neural network could only reach 46.70% and 50.57%, resp.

While the approaches mentioned above are all based on continuous HMMs, Nwe *et al.* [10] run experiments on discrete HMMs. Short acted utterances in 6 archetypal emotions obtained from 12 non-professional speakers were modelled by HMMs with up to 8 states. The choice of an ergodic topology instead of a left-right structure was deduced from the assumption that emotional cues contained in an utterance may not occur strictly sequentially. Best results were achieved for 4 states. Other approaches with discrete HMMs are reported by Pao *et al.* [11], Kang *et al.* [7] and Nogueiras *et al.* [9], all carried out on acted speech.

All these reports prove that HMMs provide a suitable method to model emotional cues in speech. If tested, results were comparable or even superior to classification based on global statistics. However, there is not yet a clear line which HMM design should be prioritised, since the used networks differ greatly in respect of the number of states, topology and model densities. If within the same approach different network configurations were compared, often only the number of states was considered, while little is known about the effect of the other parameters. On the other hand, the previous approaches are mainly based on acted speech collected from a small number of subjects, so it is still vague to what extent HMMs are also applicable to spontaneous speech.

2.3 Scope of This Work

In this work we try to overcome the before mentioned limitations by working with a considerable larger set of HMMs. Furthermore, two databases are evaluated: one containing acted speech from 10 adults, similar to the ones used in previous approaches, and a second one recorded within a more realistic setting gathering spontaneous speech from 51 children. To draw a comparison which kind of network configuration is preferable for each segmentation level, experiments are carried out on word and utterance level, respectively. In respect to previous work, we especially investigate the following assumptions:

A1: In the case of continuous models, performance is improved when the number of mixtures per state is increased, that is, detailed density modelling [13]. A comparison with discrete models has not been drawn yet.

A2: High network connectivity is beneficial for modelling the presented sequences of emotional cues [10,11].

A3: Increasing the number of states, that is, detailed temporal modelling, leads to better recognition results [8,13,6,10,9].

A4: HMM-based classifiers are less useful in applications with utterances of variable length [8].

3 Experimental Settings

3.1 Databases

Our experiments were conducted on two different databases, one with acted emotions and one with elicited spontaneous emotions.

The Berlin database of emotional speech [3] contains recordings of 10 non-professional actors (5 male/5 female) in 6 emotional states (joy, anger, boredom, disgust, sadness, fear) as well as a neutral state. There are in total 493 utterances with 4827 words in it. We used a 5-fold cross validation strategy for evaluation, with always 8 speakers in the training and 2 (1 male/1 female) in the test set. This database has been labelled on utterance level; for word level investigations, each word obtained the label of the pertaining utterance. In the following this database is referred to as BERLIN-W (words) and BERLIN-U (utterances).

The German Aibo Emotion Corpus [1] contains spontaneous emotions and has been collected in a Wizard-of-Oz setting to elicit emotions. It covers speech of 51 children (31 girls/20 boys) interacting with Sony's robot dog Aibo and was recorded at two schools. The most prominent emotional states where angry, emphatic (as a pre-stage to anger), motherese (or baby-talk) and neutral. We evaluated a subcorpus of the original corpus with a relatively balanced class distribution. Additionally to word segmentation, a segmentation into chunks exists, which have been extracted from the dialogue turns by a manually revised pause segmentation, so in total 4543 chunks and 16427 words were analysed. Labels were originally assigned to words; these were mapped onto chunks using a modified majority-voting strategy, where the impact of neutral was weakened. For a detailed description of the selection of the subcorpus and the mapping of word-based labels onto chunks or turns see [2]. This corpus was evaluated by taking one school (Ohm, 26 speakers) for training and the other (Mont, 25 speakers) for testing. In the following AIBO-W (words) and AIBO-U (utterances) are used as abbreviations for this database.

3.2 Networks

Since the number of HMM networks is theoretically infinite, first a meaningful subset of representative networks was determined.

We compared discrete and continuous networks modelled by one of the following probability distributions: discrete with size 64 or 256, and continuous consisting of 1, 4 or 8 Gaussian mixtures. As a further constraint all states of a network had to be modelled by the same number of distributions. In the following `d64`, `d256`, `c01`, `c04` and `c08` are used as abbreviations for the corresponding networks.

Next, we compared four different topologies: a linear model with only forward transition (`F`) and a left-right model (`Fj`), which are both commonly used in speech recognition, as well as a model with also backward transitions (`FB`) and a fully connected network (`FjBj`). The latter is known as an ergodic model and assumed to be notably beneficial in modelling emotional cues [10,11]. Example diagrams for each network type have been previously shown in Figure 1.

In simple word recognition tasks the number of states usually corresponds roughly to the number of phonemes within a word, that is 2 to 10 states [12]. Hence, networks with 5 and 10 states were included as a promising length for recognition on word level. Additionally — since we also try to recognize emotions based on utterance level — longer networks with 15, 20 and 25 states are

considered, as well. Networks with only a single state, which is equivalent to a GMM, were also tested in order to get an impression how much recognition gain is actually added by temporal modelling. In the following the state number of a network is coded with `sXX`.

Based on these parameters a set of 120 HMMs can be built, each coded by the concatenation of the introduced abbreviations. For example, `c04-Fj-05` stands for a continuous left-right model with 5 states and 4 Gaussian mixtures.

3.3 Feature Extraction and Evaluation

Since the modelling task was of main interest to this work, we used a relatively simple feature set consisting of 13 MFCC coefficients including the 0th cepstral parameter, which represents the energy within the frame. First and second derivation were also added, resulting in 39 features in total. MFCCs are commonly used in speech recognition to encode the spectral information of speech. In particular, they are known for their good approximation of the human auditory perception. Feature extraction, as well as evaluation was done using the Hidden Markov Model Toolkit[1] (HTK) developed by the Cambridge University Engineering Department. For each emotion we trained a separate model and classified an unknown sequence into the model that gave highest probability.

4 Results

4.1 Recognition Rates Achieved with HMMs

Each of the 120 networks was trained and tested on both databases and for both segmentation types, namely words and utterances. Sole exception: on word level results for `F` and `FB` networks with 15 or more states were discarded, since too many samples would have been refused[2]. Throughout the following chapter performance of a network is measured according to the classwise averaged recognition result CL, that is the mean of the recognition rates achieved for each emotional class.

First of all, it should be mentioned that finding general tendencies was rather difficult, since on the one hand quite different parameters sometimes gained the same results, whereas on the other hand a slight parameter change sometimes caused a very different performance. However, to get an impression of the results, Table 1 lists for each domain performance of the worst and best networks, as well as the average result of all networks. In the first place, the broad margin of at least 10 % between worst and best performing networks is remarkable. This clearly shows that the network design is a crucial aspect, which has considerable impact on recognition accuracy. Before we discuss more accurately the influence of each parameter, we also want to note that there is an obvious difference of

[1] http://htk.eng.cam.ac.uk/

[2] Networks without skipping transition can only process a sample if it has at least as many frames as the network has states.

Table 1. Worst, best and average CL in % through all networks

Words					
	worst network		best network		avg
AIBO	**43.53**	d64-FjBj-s25	**56.50**	c08-FB-s10	**52.18**
BERLIN	**33.32**	d64-Fj-s15	**50.53**	c08-FjBj-s10	**42.29**

Utterances					
	worst network		best network		avg
AIBO	**44.73**	d256-FjBj-s20	**55.80**	c04-FjBj-s15	**50.85**
BERLIN	**48.97**	d256-F-s20	**73.92**	c08-F-s05	**61.36**

performance according to the segmentation level of both databases: while average recognition results for BERLIN is about 10% better for utterances than words, it is the other way round for AIBO. An explanation for this finding will be given in the next chapter.

First, we address assumption *A1*, i. e. the impact of the probability density. Therefore, the mean CL among all networks (grey bar), as well as the average of the 3 best performing networks (white bar), is presented in Figure 2. A general improvement in performance can be observed for HMMs modelled with a continuous probability density compared to discrete HMMs, even though the difference is more obvious for BERLIN than for AIBO. It also appears that among the continuous HMMs `c04` and `c08` networks slightly outperform those models with only a single mixture, that is `c01`. However, there seems to be no general improvement when increasing the number of mixtures from 4 to 8 in continuous networks or using a larger codebook than 64 in the discrete case.

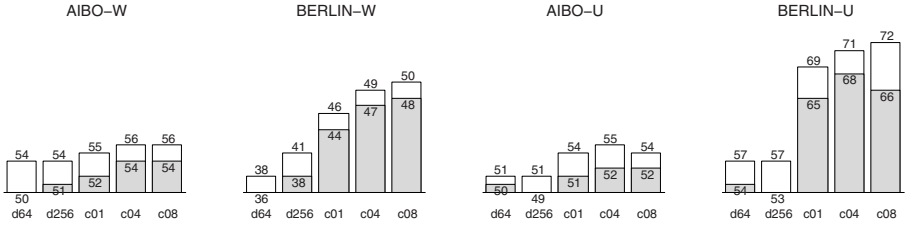

Fig. 2. Averaged CL among all networks modelled by the same probability density

Next we examine assumption *A2*, namely topology. Again, average CLs are shown in Figure 3. The results imply that none of the tested connectivity levels seem to be superior to the others. In particular, the hypothesis that additional connectivity is more suitable in modelling emotional cues is not supported.

According to assumption *A3*, average CLs according to different state numbers are given in Figure 4. Even though, differences are again small, networks of length 5 or 10 seem to be generally more profitable on word level. This supports our assumption that here similar network sizes as in word-based speech

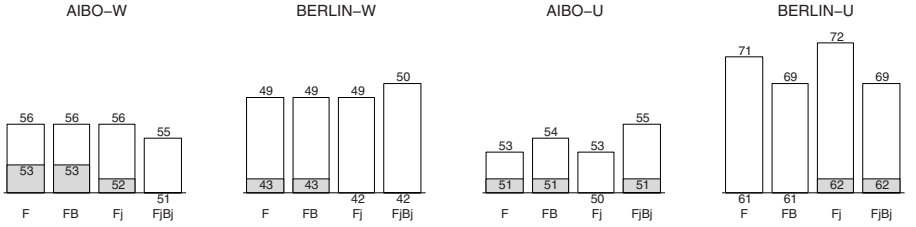

Fig. 3. Averaged CL among all networks modelled by the same topology

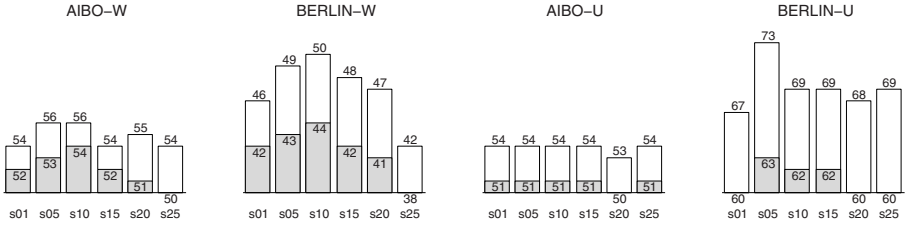

Fig. 4. Averaged CL among all networks modelled by the same number of states

recognition should be used. With regard to utterances, different tendencies for both databases can be observed: while networks with 5–15 states work best for BERLIN-U, performance for AIBO-U stays almost steady along all tested sizes. As a general tendency, results do not improve for models with 15 or more states.

To sum up, we can draw the following conclusions concerning *A1–3*:

- HMMs modelled by continuous densities outperform discrete networks.
- A pool of 4 mixtures seems to be sufficient.
- Ergodic models are not necessarily more suitable.
- HMMs with 5 to 10 states appear to be most beneficial.
- There seems to be no improvement for networks with 15 or more states.

Futhermore, results show that a good network design seems to be relatively independent of the source of speech (acted vs. spontaneous) and the segmentation level (word vs. utterance).

4.2 Interpretation of the Presented Results

As already mentioned, a remarkable difference of recognition accuracy between words and utterances can be observed for both databases. This difference follows from the different annotation strategies that have been used. While AIBO is labelled word-wise and therefore obviously adequately for evaluation on word level, words in BERLIN receive their label from the pertaining utterance. The latter certainly involves that some words do not carry the affective information they are labelled with. On the other hand, mapping labels from words to utterances — as done in AIBO — is also error prone.

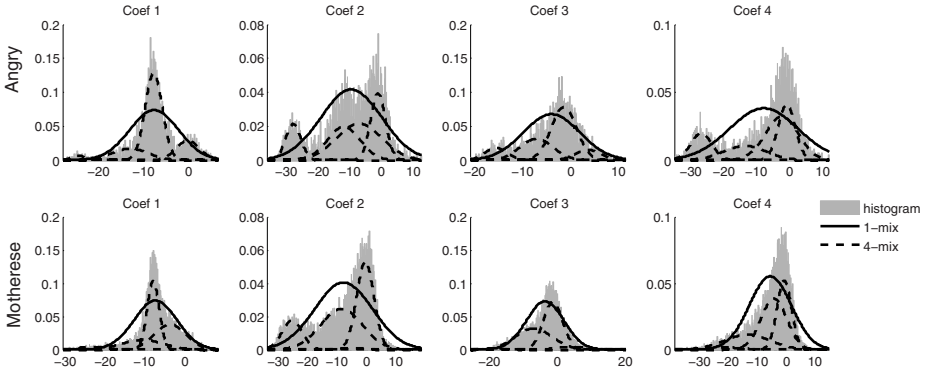

Fig. 5. Histograms and mixtures for four MFCC coefficients during motherese and angry

The output probability distribution of the HMM states defines how accurate the emotional cues are captured. Figure 5 shows histograms for the first 4 MFCC coefficients collected from all chunks of a single AIBO speaker during angry and motherese. The chart reveals that distributions belonging to the same coefficient differ only slightly between both emotions. However, a pool of 4 mixtures (dotted lines) allows a better modulation of these differences than a single Gaussian mixture (solid line). If we think of the additional distortion introduced by quantization, this also explains why discrete modelling is even less profitable.

The observation that performance of a network is more or less independent of its level of connectivity implies that additional backward and forward transitions do not significantly improve its capability to model emotional cues. Figure 6 shows transition probabilities for the `c08-FB-s05` and `c08-Fj-s05` networks trained with neutral words from AIBO-W. Indeed — compared to forward transitions — the probability associated with backward transitions is rather small in the `FB` network. Similarly, for the `Fj` network the transition probabilities become smaller the further the connected states are apart from each other. For both networks, the probability to remain in the same state is always the highest. This induces that successive frames tend to stay in the same state and that in case of a change usually the right neighbour is taken, just like we would expect it for an `F` network.

Figure 7 shows output probability densities of the 1st MFCC coefficient for states of different networks which were trained with emphatic words or utterances from AIBO. From these charts we can explain our findings according to the size of a network. For instance, we have seen that increasing the number of states to 15 or more states gave no further gain to the recognition performance. The reason is that in long networks successive states tend to show increasingly similar densities. Hence, at some point additional states do not improve temporal modelling anymore. For instance, in the network at the top of Figure 7 state 3 to 6 could be merged to a single state. In contrast, for the network shown below, which has only 5 instead of 15 states, all densities differ clearly. However, if the

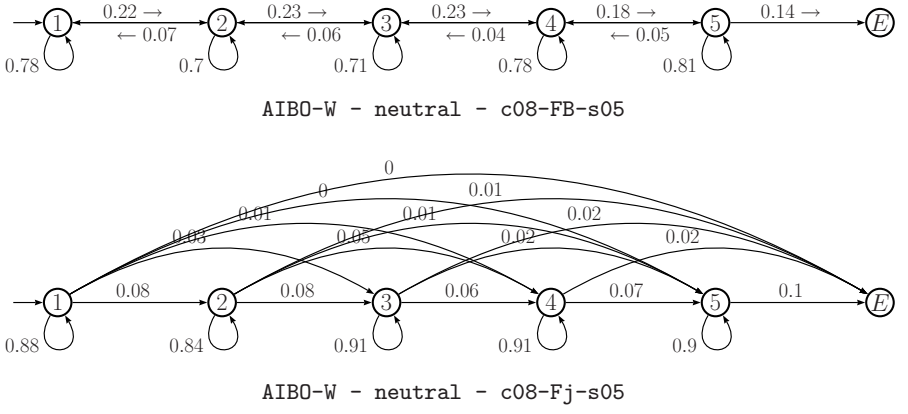

Fig. 6. Transition probabilities of a trained network with different topologies

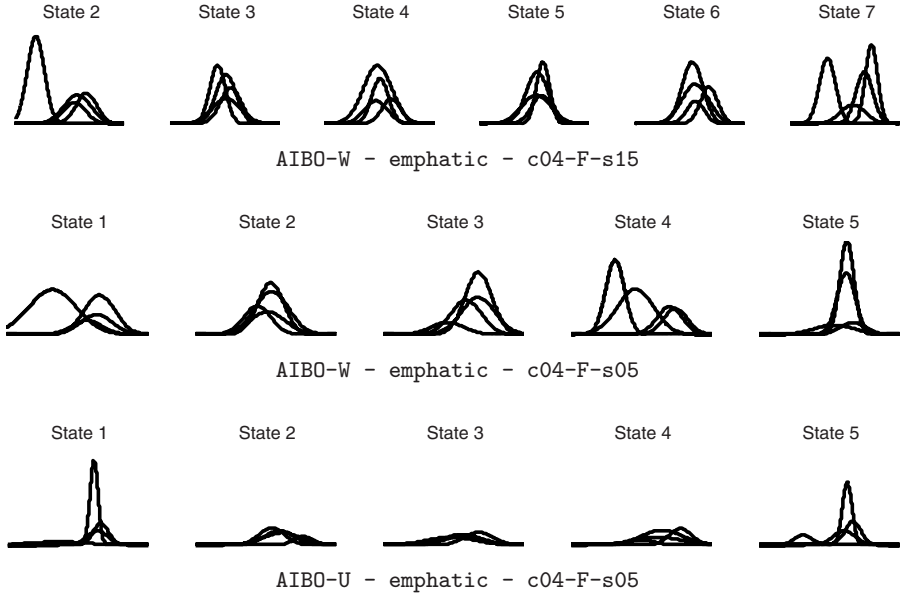

Fig. 7. Output probability densities of a single feature for states of different networks

same network is trained on utterances (bottom of Figure 7), we get again similar densities for state 2 to 4. A sign that in this case the network could be pruned even further. Indeed, this corresponds with the fact that temporal modelling did hardly gain any profit for AIBO-U. In our opinion the reason is that the utterances in this corpus range from single words to long phrases, and are therefore too inhomogeneous in respect of their length that a single network could effectively represent their temporal structure. This also explains why temporal

modelling was more successful for BERLIN-U: here utterances are exclusively whole sentences of similar length.

4.3 Comparison with a Static Classifier

For comparison purposes, we also carried out experiments with features based on global statistics. A feature set composed of 1053 MFCC and 137 energy features[3] was calculated and used to train a support vector machine (SVM). To get a fair rating for the HMM-based approach, we discarded all discrete networks, as well as continuous networks with a single Gaussian density and 15 or more states and averaged outcomes among the remaining HMMs. Results are listed in Table 2 and show that on word level — independent of the database — HMMs are more than 10 % superior to the SVM approach. If we compare the values on utterance level, HMMs are still slightly better for AIBO, but inferior for BERLIN. This supports the assumption posited by Kwon *et al.* [8] that HMM-based classifiers are less applicable in applications with utterances of variable length (*A4*).

Table 2. CL in % for the SVM and HMM-based approach

Words			Utterances		
	SVM	HMM		SVM	HMM
AIBO	**43.7**	**55.5**	AIBO	**51.2**	**52.5**
BERLIN	**36.6**	**48.6**	BERLIN	**73.3**	**69.5**

At this point we would like to mention, that performance with static modelling can be significantly improved when other feature types, such as duration or pitch features, are added. For instance, in combination with automatic feature selection, we could achieve a best recognition rate of 77.4% for BERLIN-U [14]. Using an optimised set of 381 features, involving also spectral and lexical features, Batliner *et al.* [2] report 58.7% recognition accuracy for turns in AIBO.

5 Conclusion

Our results confirm that HMMs provide a suitable method to model emotional cues in speech. Compared to classification based on global statistics, performance was similar on utterance level and clearly superior on word level. This applied for acted and spontaneous speech. From this we conclude that HMMs are also applicable in realistic settings.

Our results also show the difficulty to set up general guidelines which kind of networks is best suited. Tests based on a large set of networks could, for instance, not support the assumption that ergodic models are generally more suitable in modelling emotional cues. Solely, continuous networks gave significantly better results compared to discrete HMMs. According to the source of speech (acted vs.

[3] To achieve comparable results we obtained global statistics only for those feature types also used in our approach with HMMs.

spontaneous) and the segmentation level (word vs. utterance) similar tendencies were observed.

In [8], Kwon *et al.* state that HMM-based classifiers are less applicable for utterances of variable length. Our experiments prove this assumption insofar as temporal modelling appeared to be less profitable when the modelled samples were very diverse in respect of their length.

In our future work we will investigate this assumption more accurately by using a pool of networks which are trained with samples of similar sizes. Likewise, multiple networks can be used to represent nuances of the same emotion, such as cold and hot anger. Furthermore, it seems reasonable to combine dynamic and static classification to get the benefits of both approaches.

Acknowledgements. This work was partially funded by the EU network of excellence HUMAINE and the EU projects eCIRCUS and CALLAS.

References

1. Batliner, A., Hacker, C., Steidl, S., Nöth, E., D'Arcy, S., Russell, M., Wong, M.: You stupid tin box — children interacting with the AIBO robot: A cross-linguistic emotional speech corpus, LREC, Lisbon, Portugal (2004)
2. Batliner, A., Steidl, S., Schuller, B., Seppi, D., Laskowski, K., Vogt, T., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., Aharonson, V.: Combining Efforts for Improving Automatic Classification of Emotional User States, IS-LTC, Ljubljana, Slov (2006)
3. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F., Weiss, B.: A Database of German Emotional Speech, Interspeech, Lisbon, Portugal (2005)
4. Cowie, R., Douglas-Cowie, E., Tsapatoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.G.: Emotion Recognition in Human-Computer Interaction. IEEE Signal Processing Magazine 18(1), 32–80 (2001)
5. Fernandez, R., Picard, R.W.: Modeling drivers' speech under stress. Speech Communication 40(1-2), 145–159 (2003)
6. Jiang, D.-N., Cai, L.-H.: Speech emotion classification with the combination of statistic features and temporal features, ICME, Taipei, Taiwan (2004)
7. Kang, B.-S., Han, C.-H., Lee, S.-T., Youn, D.-H., Lee, C.: Speaker dependent emotion recognition using speech signals, ICSLP, Beijing, China (2000)
8. Kwon, O.-W., Chan, K.-L., Hao, J., Lee, T.-W.: Emotion Recognition by Speech Signals, Eurospeech, Geneva, Switzerland (2003)
9. Nogueiras, A., Moreno, A., Bonafonte, A., Mari, J.B.: Speech emotion recognition using hidden Markov models, Eurospeech, Aalborg, Denmark (2001)
10. Nwe, T.L., Foo, S.W., De Silva, L.C.: Speech emotion recognition using hidden Markov models. Speech Communication 41(4), 603–623 (2003)
11. Pao, T.-L., Chen, Y.-T., Yeh, J.-H., Liao, W.-Y.: Detecting Emotions in Mandarin Speech. Comp. Ling. and Chinese Lang. Proc. 10(3), 347–362 (2005)
12. Rabiner, L.R.: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proc. IEEE 77(2), 257–286 (1989)
13. Schuller, B., Rigoll, G., Lang, M.: Hidden Markov model-based speech emotion recognition, ICME, Baltimore, USA (2003)
14. Vogt, T., André, E.: Comparing Feature Sets for Acted and Spontaneous Speech in View of Automatic Emotion Recognition. ICME (2005)